BioNLP Shared Task 2013: Supporting Resources

Pontus Stenetorp [1] **Wiktoria Golik** [2] **Thierry Hamon** [3]
Donald C. Comeau [4] **Rezarta Islamaj Doğan** [4] **Haibin Liu** [4] **W. John Wilbur** [4]
[1] National Institute of Informatics, Tokyo, Japan
[2] French National Institute for Agricultural Research (INRA), Jouy-en-Josas, France
[3] University Paris 13, Paris, France
[4] National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD, USA
pontus@nii.ac.jp wiktoria.golik@jouy.inra.fr thierry.hamon@univ-paris13.fr
{comeau,islamaj,liuh11,wilbur}@ncbi.nlm.nih.gov

Abstract

This paper describes the technical contribution of the supporting resources provided for the BioNLP Shared Task 2013. Following the tradition of the previous two BioNLP Shared Task events, the task organisers and several external groups sought to make system development easier for the task participants by providing automatically generated analyses using a variety of automated tools. Providing analyses created by different tools that address the same task also enables extrinsic evaluation of the tools through the evaluation of their contributions to the event extraction task. Such evaluation can improve understanding of the applicability and benefits of specific tools and representations. The supporting resources described in this paper will continue to be publicly available from the shared task homepage http://2013.bionlp-st.org/

1 Introduction

The BioNLP Shared Task (ST), first organised in 2009, is an ongoing series of events focusing on novel challenges in biomedical domain information extraction. In the first BioNLP ST, the organisers provided the participants with automatically generated syntactic analyses from a variety of Natural Language Processing (NLP) tools (Kim et al., 2009) and similar syntactic analyses have since then been a key component of the best performing systems participating in the shared tasks. This initial work was followed up by a similar effort in the second event in the series (Kim et al., 2011), extended by the inclusion of software tools and contributions from the broader BioNLP community in addition to task organisers (Stenetorp et al., 2011).

Although no formal study was carried out to estimate the extent to which the participants utilised the supporting resources in these previous events, we note that six participating groups mention using the supporting resources in published descriptions of their methods (Emadzadeh et al., 2011; McClosky et al., 2011; McGrath et al., 2011; Nguyen and Tsuruoka, 2011; Björne et al., 2012; Vlachos and Craven, 2012). These resources have been available also after the original tasks, and several subsequent studies have also built on the resources. Van Landeghem et al. (2012) applied a visualisation tool that was made available as a part of the supporting resources, Vlachos (2012) employed the syntactic parses in a follow-up study on event extraction, Van Landeghem et al. (2013) used the parsing pipeline created to produce the syntactic analyses, and Stenetorp et al. (2012) presented a study of the compatibility of two different representations for negation and speculation annotation included in the data.

These research contributions and the overall positive reception of the supporting resources prompted us to continue to provide supporting resources for the BioNLP Shared Task 2013. This paper presents the details of this technical contribution.

2 Organisation

Following the practice established in the BioNLP ST 2011, the organisers issued an open call for supporting resources, welcoming contributions relevant to the task from all authors of NLP tools. In the call it was mentioned that points such as availability for research purposes, support for well-established formats and access

Proceedings of the BioNLP Shared Task 2013 Workshop, pages 99–103,
Sofia, Bulgaria, August 9 2013.

Name	Annotations	Availability
BioC	Lemmas and syntactic constituents	Source
BioYaTeA	Terms, lemmas, part-of-speech and syntactic constituencies	Source
Cocoa	Entities	Web API

Table 1: Summary of tools/analyses provided by external groups.

to technical documentation would considered favourable (but not required) and each supporting resource provider was asked to write a brief description of their tools and how they could potentially be applied to aid other systems in the event extraction task. This call was answered by three research groups that offered to provide a variety of semantic and syntactic analyses. These analyses were provided to the shared task participants along with additional syntactic analyses created by the organisers.

However, some of the supporting resource providers were also participants in the main event extraction tasks, and giving them advance access to the annotated texts for the purpose of creating the contributed analyses could have given those groups an advantage over others. To address this issue, the texts were made publicly available one week prior to the release of the annotations for each set of texts. During this week, the supporting analysis providers annotated the texts using their automated tools and then handed the analyses to the shared task organisers, who made them available to the task participants via the shared task homepage.

3 Analyses by External Groups

This section describes the tools that were applied to create supporting resources by the three external groups. These contributions are summarised in Table 1.

BioC Don Comeau, Rezarta Islamaj, Haibin Liu and John Wilbur of the National Center for Biotechnology Information provided the output of the shallow parser MedPost (Smith et al., 2004) and the BioLemmatizer tool (Liu et al., 2012), supplied in the BioC XML format[1] for annotation interchange (Comeau et al., 2013). The BioC format address the problem of interoperability between different tools and platforms by providing a unified format for use by various tools. Both MedPost and BioLemmatizer are specifically designed for biomedical texts. The former annotates parts-of-speech and performs sentence splitting and tokenisation, while the latter performs lemmatisation. In order to make it easier for participants to get started with the BioC XML format, the providers also supplied example code for parsing the format in both the Java and C++ programming languages.

BioYaTeA Wiktoria Golik of the French National Institute for Agricultural Research (INRA) and Thierry Hamon of University Paris 13 provided analyses created by BioYaTeA[2] (Golik et al., 2013). BioYaTeA is a modified version of the YaTeA term extraction tool (Aubin and Hamon, 2006) adapted to the biomedical domain. Working on a noun-phrase level, BioYaTeA provides annotations such as lemmas, parts-of-speech, and constituent analysis. The output formats used were a simple tabular format as well as BioYaTeA-XML, an XML representation specific to the tool.

Cocoa S. V. Ramanan of RelAgent Private Ltd provided the output of the Compact cover annotator (Cocoa) for biological noun phrases.[3] Cocoa provides noun phrase-level entity annotations for over 20 different semantic categories such as macromolecules, chemicals, proteins and organisms. These annotations were made available for the annotated texts for the shared task along with the opportunity for the participants to use the Cocoa web API to annotate any text they may consider beneficial for their system. The data format used by Cocoa is a subset of the standoff format used for the shared task entity annotations, and it should thus be easy to integrate into existing event extraction systems.

4 Analyses by Task Organisers

This section describes the syntactic parsers applied by the task organisers and the pre-processing

[1] http://bioc.sourceforge.net/

[2] http://search.cpan.org/~bibliome/Lingua-BioYaTeA/

[3] http://npjoint.com/

Name	Model	Availability
Enju	Biomedical	Binary
Stanford	Combination	Binary, Source
McCCJ	Biomedical	Source

Table 2: Parsers used for the syntactic analyses.

and format conversions applied to their output. The applied parsers are listed in Table 2.

4.1 Syntactic Parsers

Enju Enju (Miyao and Tsujii, 2008) is a deep parser based on the Head-Driven Phrase Structure Grammar (HPSG) formalism. Enju analyses its input in terms of phrase structure trees with predicate-argument structure links, represented in a specialised XML-format. To make the analyses of the parser more accessible to participants, we converted its output into the Penn Treebank (PTB) format using tools included with the parser. The use of the PTB format also allow for its output to be exchanged freely for that of the other two syntactic parsers and facilitates further conversions into dependency representations.

McCCJ The BLLIP Parser (Charniak and Johnson, 2005), also variously known as the Charniak parser, the Charniak-Johnson parser, or the Brown reranking parser, has been applied in numerous biomedical domain NLP efforts, frequently using the self-trained biomedical model of McClosky (2010) (i.e. the McClosky-Charniak-Johnson or McCCJ parser). The BLLIP Parser is a constituency (phrase structure) parser and the applied model produces PTB analyses as its native output. These analyses were made available to participants without modification.

Stanford The Stanford Parser (Klein and Manning, 2002) is a widely used publicly available syntactic parser. As for the Enju and BLLIP parsers, a model trained on a dataset incorporating biomedical domain annotations is available also for the Stanford parser. Like the BLLIP parser, the Stanford parser is constituency-based and produces PTB analyses, which were provided to task participants. The Stanford tools additionally incorporate methods for automatic conversion from this format to other representations, discussed further below.

4.2 Pre-processing and Conversions

To create the syntactic analyses from the Enju, BLLIP and Stanford Parser systems, we first applied a uniform set of pre-processing steps in order to normalise over differences in e.g. tokenisation and thus ensure that the task participants can easily swap the output of one system for another. This pre-processing was identical to that applied in the BioNLP 2011 Shared Task, and included sentence splitting of the annotated texts using the Genia Sentence Splitter,[4] the application of a set of post-processing heuristics to correct frequently occurring sentence splitting errors, and Genia Treebank-like tokenisation (Tateisi et al., 2004) using a tokenisation script created by the shared task organisers. [5]

Since several studies have indicated that representations of syntax and aspects of syntactic dependency formalism differ in their applicability to support information extraction tasks (Buyko and Hahn, 2010; Miwa et al., 2010; Quirk et al., 2011), we further converted the output of each of the parsers from the PTB representation into three other representations: CoNNL-X, Stanford Dependencies and Stanford Collapsed Dependencies. For the CoNLL-X format we employed the conversion tool of Johansson and Nugues (2007), and for the two Stanford Dependency variants we used the converter provided with the Stanford CoreNLP tools (de Marneffe et al., 2006). These analyses were provided to participants in the output formats created by the respective tools, i.e. the TAB-separated column-oriented format CoNLL and the custom text-based format of the Stanford Dependencies.

5 Results and Discussion

Just like in previous years the supporting resources were well-received by the shared task participants and as many as five participating teams mentioned utilising the supporting resources in their initial submissions (at the time of writing, the camera-ready versions were not yet available). This level of usage of the supporting resources by the participants is thus comparable to what was observed for the 2011 shared task.

Following in the tradition of the 2011 support-

[4] https://github.com/ninjin/geniass

[5] https://github.com/ninjin/bionlp_st_2013_supporting/blob/master/tls/GTB-tokenize.pl

ing resources, to aim for reproducibility, the processing pipeline containing pre/post-processing and conversion scripts for all the syntactic parses has been made publicly available under an open licence.[6] The repository containing the pipeline also contains detailed instructions on how to reproduce the output and how it can potentially be applied to other texts.

Given the experience of the organisers in analysing medium-sized corpora with a variety of syntactic parsers, many applied repeatedly over several years, we are also happy to report that the robustness of several publicly available parsers has recently improved noticeably. Random crashes, corrupt outputs and similar failures appear to be transitioning from being expected to rare occurrences.

In this paper, we have introduced the supporting resources provided for the BioNLP 2013 Shared Task by the task organisers and external groups. These resources included both syntactic and semantic annotations and were provided to allow the participants to focus on the various novel challenges of constructing event extraction systems by minimizing the need for each group to separately perform standard processing steps such as syntactic analysis.

Acknowledgements

We would like to give special thanks to Richard Johansson for providing and allowing us to distribute an improved and updated version of his format conversion tool.[7] We would also like to express our appreciation to the broader NLP community for their continued efforts to improve the availability of both code and data, thus enabling other researchers to stand on the shoulders of giants.

This work was partially supported by the Quaero programme funded by OSEO (the French agency for innovation). The research of Donald C. Comeau, Rezarta Islamaj Doğan, Haibin Liu and W. John Wilbur was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

[6] https://github.com/ninjin/bionlp_st_2013_supporting

[7] https://github.com/ninjin/pennconverter

References

Sophie Aubin and Thierry Hamon. 2006. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, 13(Suppl 11):S4.

Ekaterina Buyko and Udo Hahn. 2010. Evaluating the Impact of Alternative Dependency Graph Encodings on Solving Event Extraction Tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 982–992, Cambridge, MA, October.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics.

Donald C. Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. 2013. BioC: A minimalist approach to interoperability for biomedical text processing. submitted.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Ehsan Emadzadeh, Azadeh Nikfarjam, and Graciela Gonzalez. 2011. Double layered learning for biological event extraction from text. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 153–154. Association for Computational Linguistics.

Wiktoria Golik, Robert Bossy, Zorana Ratkovic, and Claire Nédellec. 2013. Improving Term Extraction with Linguistic Analysis in the Biomedical Domain. In *Special Issue of the journal Research in Computing Science*, Samos, Greece, March. 14th International Conference on Intelligent Text Processing and Computational Linguistics.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proc. of the 16th Nordic Conference on Computational Linguistics (NODALIDA)*, pages 105–112.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of Genia Event Task in BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*.

Dan Klein and Christopher D Manning. 2002. Fast exact inference with a factored model for natural language parsing. *Advances in neural information processing systems*, 15(2003):3–10.

Haibin Liu, Tom Christiansen, William Baumgartner, and Karin Verspoor. 2012. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3(1):3.

David McClosky, Mihai Surdeanu, and Christopher D Manning. 2011. Event Extraction as Dependency Parsing for BioNLP 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 41–45. Association for Computational Linguistics.

David McClosky. 2010. *Any domain parsing: Automatic domain adaptation for natural language parsing*. Ph.D. thesis, Brown University.

Liam R McGrath, Kelly Domico, Courtney D Corley, and Bobbie-Jo Webb-Robertson. 2011. Complex biological event extraction from full text using signatures of linguistic and semantic features. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 130–137. Association for Computational Linguistics.

Makoto Miwa, Sampo Pyysalo, Tadayoshi Hara, and Jun'ichi Tsujii. 2010. Evaluating Dependency Representations for Event Extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 779–787, Beijing, China, August.

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80.

Nhung TH Nguyen and Yoshimasa Tsuruoka. 2011. Extracting bacteria biotopes with semi-supervised named entity recognition and coreference resolution. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 94–101. Association for Computational Linguistics.

Chris Quirk, Pallavi Choudhury, Michael Gamon, and Lucy Vanderwende. 2011. MSR-NLP Entry in BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 155–163, Portland, Oregon, USA, June.

Larry Smith, Thomas Rindflesch, and W. John Wilbur. 2004. MedPost: a part-of-speech tagger for bio medical text. *Bioinformatics*, 20(14):2320–2321.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. BioNLP Shared Task 2011: Supporting Resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 112–120, Portland, Oregon, USA, June.

Pontus Stenetorp, Sampo Pyysalo, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Bridging the gap between scope-based and event-based negation/speculation annotations: a bridge not too far. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 47–56. Association for Computational Linguistics.

Y Tateisi, T Ohta, and J Tsujii. 2004. Annotation of predicate-argument structure on molecular biology text. *Proceedings of the Workshop on the 1st International Joint Conference on Natural Language Processing (IJCNLP-04)*.

Sofie Van Landeghem, Kai Hakala, Samuel Rönnqvist, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2012. Exploring biomolecular literature with EVEX: connecting genes through events, homology, and indirect associations. *Advances in Bioinformatics*, 2012.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, et al. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PloS one*, 8(4):e55814.

Andreas Vlachos and Mark Craven. 2012. Biomedical event extraction from abstracts and full papers using search-based structured prediction. *BMC Bioinformatics*, 13(Suppl 11):S5.

Andreas Vlachos. 2012. An investigation of imitation learning algorithms for structured prediction. In *Workshop on Reinforcement Learning*, page 143.